COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, Image, Vision*

Par
Kevin Riou

Embodiment- and Environment-Agnostic Imitation Learning for Robots : Integrating Human Pose-Based Action Recognition with Language and Vision Models.

Thèse présentée et soutenue à Nantes, le
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Catherine ACHARD	Professeur des universités, Sorbonne Université, France
Vitor SANTOS	Associate professor with Habilitation, University of Aveiro, Portugal

Composition du Jury :

Président :	Marie BABEL	Professeur des universités, INSA de Rennes, France
Examinateurs :	Catherine ACHARD	Professeur des universités, Sorbonne Université, France
	Vitor SANTOS	Associate professor, University of Aveiro, Portugal
	Diana MATEUS	Professeur des universités, Centrales Nantes, France
	Youcef MEZOUAR	Professeur des universités, Université Clermont Auvergne, France
Dir. de thèse :	Patrick LE CALLET	Professeur des universités, Nantes Université, France
Co-dir. de thèse :	Kevin SUBRIN	Maitre de conférence, Nantes Université, France

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : Apprentissage par Imitation Indépendant de l'Incarnation et de l'Environnement pour les Robots : Intégration de la Reconnaissance d'Actions Basée sur la Pose Humaine avec des Modèles de Langage et Vision.

Mot clés : Apprentissage par Imitation, Estimation de Pose Humaine, Représentation de Scène, Robotique

Résumé : L'apprentissage automatique, notamment l'apprentissage par imitation (IL), permet aux robots de gérer des tâches complexes dans des environnements non structurés. De plus, l'IL permet aux robots d'apprendre de nouvelles tâches à partir de démonstrations humaines sans programmation manuelle, mais les méthodes actuelles nécessitent de nombreuses démonstrations téléopérées, ce qui est intrusif, chronophage et limitant pour l'expert qui réalise la démonstration. De plus, les politiques de contrôle robotique apprises à partir de ces données sont généralement spécifiques au robot et à l'environnement illustrés dans les démonstrations. Cette thèse propose d'entraîner des politiques de contrôle directement à partir de vidéos d'humains exécutant des tâches, en se concentrant sur trois axes principaux : (1) utiliser des modèles de langage et de vision pour identifier les objets d'intérêt dans la scène et créer une représentation 3D indépendante de l'environnement de fond et de l'agent exécutant la tâche (humain/robot), (2) appliquer l'estimation de la pose humaine 3D pour extraire les actions humaines des démonstrations vidéo, et (3) entraîner une politique en IL pour prédire ces actions à partir de la représentation 3D proposée, et valider ses capacités de généralisation à de nouveaux agents et environnements en la déployant sur un robot simulé.

Title: Embodiment- and Environment-Agnostic Imitation Learning for Robots: Integrating Human Pose-Based Action Recognition with Language and Vision Models.

Keywords: Imitation Learning, Human Pose Estimation, Scene Representation, Robotics

Abstract: Machine Learning, especially Imitation Learning (IL), enables robots to handle complex tasks in unstructured environments. Moreover, IL allows robots to learn new tasks from human demonstrations without manual programming, but current methods require many teleoperated demonstrations, which are intrusive, time-consuming, and limiting for the expert demonstrating the task. Additionnaly, the robot control policies learnt from such data are usually specific to the robot's hardware and environment highlighted in the demonstrations. This thesis proposes to train robot control policies directly from videos of humans performing tasks, focusing on three key areas: (1) using language and vision models to identify objects of interest and create a 3D scene representation agnostic to the background environment and to the embodiment (human/robot), (2) applying 3D Human Pose Estimation (3DHPE) to extract human actions from the video demonstrations, and (3) training an IL policy to predict these actions from the proposed 3D scene representation, and validate its generalization ability to new embodiments and environments by deploying it on a simulated robot.